(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-139043
(P2018-139043A)
(43)公開日　平成30年9月6日(2018.9.6)

(51)Int.Cl.			F I			テーマコード（参考）
G06F	***19/24***	***(2011.01)***	G06F	19/24		4B063
C12N	*15/09*	*(2006.01)*	C12N	15/00	A	
C12Q	*1/68*	*(2018.01)*	C12Q	1/68	Z	

審査請求 未請求　請求項の数 8　ＯＬ　（全 17 頁）

(21)出願番号　特願2017-33381(P2017-33381)
(22)出願日　平成29年2月24日(2017.2.24)

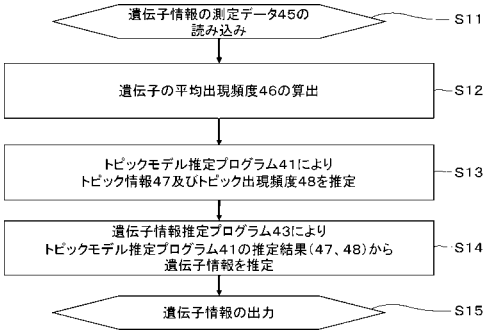
（出願人による申告）（１）平成２７年度、国立研究開発法人科学技術振興機構、戦略的創造研究推進事業「野外環境と超並列高度制御環境の統合モデリングによる頑健性限界の解明と応用」委託研究、産業技術力強化法第１９条の適用を受ける特許出願（２）平成２７年度、国立研究開発法人科学技術振興機構、戦略的創造研究推進事業「共生ネットワークの分子基盤とその応用展開」委託研究、産業技術力強化法第１９条の適用を受ける特許出願

(71)出願人　597065329
学校法人　龍谷大学
京都府京都市伏見区深草塚本町６７番地
(74)代理人　110000914
特許業務法人　安富国際特許事務所
(72)発明者　岩山　幸治
滋賀県大津市瀬田大江町横谷１－５　学校法人　龍谷大学内
(72)発明者　永野　惇
滋賀県大津市瀬田大江町横谷１－５　学校法人　龍谷大学内

Ｆターム(参考)　4B063 QA01　QA13　QQ42　QS39

(54)【発明の名称】トピックモデルを用いた遺伝子情報推定装置

(57)【要約】　（修正有）
【課題】低コストかつ高精度で遺伝子情報を推定できる装置を提供する。
【解決手段】既存のトピックモデルに基づき、遺伝子情報に適した新しいトピックモデルを用いた遺伝子情報を推定する。このモデルでは単語の生成分布を従来の多項分布から負の二項分布に置き換える。
【選択図】図２



【特許請求の範囲】
【請求項１】
（ｉ）複数のサンプルに対応する遺伝子情報の測定データを読み込む手順、
（ｉｉ）遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出する手順、
（ｉｉｉ）各サンプルの各遺伝子にトピックが割り当てられる確率を初期化する手順、
（ｉｖ）遺伝子へのトピック割り当て確率を更新する手順、
（ｖ）トピック情報を更新する手順、
（ｖｉ）トピック情報の分散を更新する手順、
（ｖｉｉ）負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新する手順、
および、
（ｖｉｉｉ）トピック情報およびトピック出現頻度から遺伝子情報を推定する手順、
を有する、
トピックモデルを用いた遺伝子情報推定装置。
【請求項２】
手順（ｉｖ）、（ｖ）、および（ｖｉｉ）において、過分散の程度を表すパラメータをϕ_iとし、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数をμ_{ski}としたとき、サンプル（ｓ）における遺伝子（ｉ）の出現数（r_{si}）が、
【数１】

$$P(r_{si}) = \frac{\Gamma\left(r_{si} + \phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)} \left(\frac{\mu_{ski}}{\mu_{ski} + \phi_i^{-1}}\right)^{r_{si}} \left(\frac{1}{1 + \mu_{ski}\phi_i}\right)^{\phi_i^{-1}}$$

という確率に従う、請求項１に記載の装置。
【請求項３】
サンプル（ｓ）における全遺伝子のカウントの総和（ν_s）、全サンプルでの全遺伝子の出現数の総和に対する遺伝子（ｉ）の出現数の割合の対数（m_i）、指数分布あるいはジェフリーズ事前分布に従う分散（τ_{ki}）、および平均が０で分散がτ_{ki}の正規分布に従うトピック（ｋ）における遺伝子（ｉ）の期待カウントの平均からのずれを表す量が与えられたときに、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数（μ_{ski}）が
$\mu_{ski} = \nu_s \exp(\eta_{ki} + m_i)$
と表される、請求項２に記載の装置。
【請求項４】
請求項１～３のいずれか一項に記載の遺伝子情報推定装置を構成する各手段としてコンピュータを機能させるための、遺伝子情報推定プログラム。
【請求項５】
請求項４に記載の遺伝子情報推定プログラムが記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項６】
（ｉ）複数のサンプルに対応する遺伝子情報の測定データを読み込むステップ、
（ｉｉ）遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出するステップ、
（ｉｉｉ）各サンプルの各遺伝子にトピックが割り当てられる確率を初期化するステップ、
（ｉｖ）遺伝子へのトピック割り当て確率を更新するステップ、
（ｖ）トピック情報を更新するステップ、
（ｖｉ）トピック情報の分散を更新するステップ、
（ｖｉｉ）負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新するステップ、
および、
（ｖｉｉｉ）トピック情報およびトピック出現頻度から遺伝子情報を推定するステップ、
を有する、トピックモデルを用いた遺伝子情報推定方法。

【請求項７】
ステップ（ｉｖ）、（ｖ）、および（ｖｉｉ）において、過分散の程度を表すパラメータをϕ_iとし、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数をμ_{ski}としたとき、サンプル（ｓ）における遺伝子（ｉ）の出現数（r_{si}）が、
【数２】

$$P(r_{si}) = \frac{\Gamma\left(r_{si} + \phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)} \left(\frac{\mu_{ski}}{\mu_{ski} + \phi_i^{-1}}\right)^{r_{si}} \left(\frac{1}{1 + \mu_{ski}\phi_i}\right)^{\phi_i^{-1}}$$

という確率に従う、請求項６に記載の方法。
【請求項８】
サンプル（ｓ）における全遺伝子のカウントの総和（ν_s）、全サンプルでの全遺伝子の出現数の総和に対する遺伝子（ｉ）の出現数の割合の対数（m_i）、指数分布あるいはジェフリーズ事前分布に従う分散（τ_{ki}）、および平均が０で分散がτ_{ki}の正規分布に従うトピック（ｋ）における遺伝子（ｉ）の期待カウントの平均からのずれを表す量が与えられたときに、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数（μ_{ski}）が
$\mu_{ski} = \nu_s \exp(\eta_{ki} + m_i)$
と表される、請求項７に記載の方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、トピックモデルを用いた遺伝子情報推定装置に関する。
【背景技術】
【０００２】
近年、核酸塩基配列を次世代シーケンサで決定することで、遺伝子情報を網羅的に解析する技術が開発されている。例えば、ＤＮＡ上の遺伝子は、転写、翻訳などを経てその機能を発現するが、この遺伝子発現の状態を解析する目的で、転写産物であるＲＮＡの配列を次世代シーケンサで決定して遺伝子の発現を網羅的に定量するＲＮＡ－Ｓｅｑ法が用いられている（非特許文献１）。
【０００３】
ＲＮＡ－Ｓｅｑ等の遺伝子情報には、数千から数万の遺伝子の情報が含まれるため、出力されるデータは高次元のカウントデータとなる。同様の高次元カウントデータを扱うことの多い自然言語処理の分野では、文書の潜在的な意味を扱う手法としてトピックモデルが提案されている。トピックモデルは、文書内の単語の共起に基づき、文書を構成するトピックと各トピックにおける単語の出現頻度を同時に推定する。トピックモデルをＲＮＡ－Ｓｅｑで得られた遺伝子発現データに適用することで、類似した発現パターンを持つ遺伝子群をトピックとして抽出できるのではないかと考えられる。実際、トピックモデルの一つであるＬａｔｅｎｔ　Ｄｉｒｉｃｈｌｅ　Ａｌｌｏｃａｔｉｏｎ（ＬＤＡ；非特許文献２）を、単一細胞のＲＮＡ－Ｓｅｑデータに適用することで、推定されたトピックから細胞間の階層的な構造を推定できることが示されている（非特許文献３）。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】Ｗａｎｇ　ｅｔ　ａｌ．Ｎａｔｕｒｅ　Ｒｅｖｉｅｗｓ　Ｇｅｎｅｔｉｃｓ　２００９　Ｊａｎ；１０（１）：５７
【非特許文献２】Ｂｌｅｉ　ｅｔ　ａｌ．Ｊｏｕｒｎａｌ　ｏｆ　Ｍａｃｈｉｎｅ　Ｌｅａｒｎｉｎｇ　Ｒｅｓｅａｒｃｈ　２００３　３：９９３
【非特許文献３】ＤｕＶｅｒｌｅ　ｅｔ　ａｌ．ＢＭＣ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ

２０１６　Ｓｅｐ　１３；１７（１）：３６３
【発明の概要】
【発明が解決しようとする課題】
【０００５】
ＲＮＡ－Ｓｅｑ等の遺伝子情報においては、読んだ配列の数（総リード数；ｄｅｐｔｈ）が多いほどノイズの少ない質の高いデータが得られる反面、コストが上がるために扱えるサンプルの数が減るというトレードオフが存在する。総リード数の少ないデータから精度よく発現量を推定できれば、より多くのサンプルの遺伝子情報発現量を定量することができる。ＲＮＡ－Ｓｅｑの場合、実際の発現データを見ると、発現パターンの相関した遺伝子が多数存在する。こうした遺伝子間の発現パターンの関係性をトピックモデルにより抽出することで、ノイズの大きな定量データからでも他の遺伝子の情報を利用して高精度な発現量の推定が可能になると期待される。しかし、ＲＮＡ－Ｓｅｑデータには、過分散、つまりポアソン分布や二項分布で予想される分散よりも大きな分散を持つという性質がある。従来のトピックモデルでは、多項分布を用いているため、このような大きな分散を説明することができない。
【課題を解決するための手段】
【０００６】
本発明では、既存のトピックモデルに基づき、過分散を有する遺伝子情報の測定データにも適用可能な新しいトピックモデル及びその推定方法を提案する。このモデルでは、従来、単語の生成分布に用いられていた多項分布を、負の二項分布に置き換える。
【０００７】
すなわち、本発明は、
（ｉ）複数のサンプルに対応する遺伝子情報の測定データを読み込む手順、
（ｉｉ）遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出する手順、
（ｉｉｉ）各サンプルの各遺伝子にトピックが割り当てられる確率を初期化する手順、
（ｉｖ）遺伝子へのトピック割り当て確率を更新する手順、
（ｖ）トピック情報を更新する手順、
（ｖｉ）トピック情報の分散を更新する手順、
（ｖｉｉ）負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新する手順、
および、
（ｖｉｉｉ）トピック情報およびトピック出現頻度から遺伝子情報を推定する手順、
を有する、
トピックモデルを用いた遺伝子情報推定装置に関する。
【０００８】
手順（ｉｖ）、（ｖ）、および（ｖｉｉ）において、過分散の程度を表すパラメータをϕ_iとし、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数をμ_{ski}としたとき、サンプル（ｓ）における遺伝子（ｉ）の出現数（r_{si}）が、
【数１】

$$P(r_{si}) = \frac{\Gamma\left(r_{si} + \phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)} \left(\frac{\mu_{ski}}{\mu_{ski} + \phi_i^{-1}}\right)^{r_{si}} \left(\frac{1}{1 + \mu_{ski}\phi_i}\right)^{\phi_i^{-1}}$$

という確率に従うことが好ましい。
【０００９】
サンプル（ｓ）における全遺伝子のカウントの総和（ν_s）、全サンプルでの全遺伝子の出現数の総和に対する遺伝子（ｉ）の出現数の割合の対数（m_i）、指数分布あるいはジェフリーズ事前分布に従う分散（τ_{ki}）、および平均が０で分散がτ_{ki}の正規分布に従うトピック（ｋ）における遺伝子（ｉ）の期待カウントの平均からのずれを表す量が与えられたときに、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数（μ_{ski}）が

$\mu_{ski} = \nu_s \exp(\eta_{ki} + m_i)$
と表される。
【００１０】
また、本発明は、前記遺伝子情報推定装置を構成する各手段としてコンピュータを機能させるための、遺伝子情報推定プログラムに関する。
【００１１】
また、本発明は、前記遺伝子情報推定プログラムが記録されたことを特徴とするコンピュータ読み取り可能な記録媒体に関する。
【００１２】
また、本発明は、
（ｉ）複数のサンプルに対応する遺伝子情報の測定データを読み込むステップ、
（ｉｉ）遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出するステップ、
（ｉｉｉ）各サンプルの各遺伝子にトピックが割り当てられる確率を初期化するステップ、
（ｉｖ）遺伝子へのトピック割り当て確率を更新するステップ、
（ｖ）トピック情報を更新するステップ、
（ｖｉ）トピック情報の分散を更新するステップ、
（ｖｉｉ）負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新するステップ、
および、
（ｖｉｉｉ）トピック情報およびトピック出現頻度から遺伝子情報を推定するステップ、
を有する、トピックモデルを用いた遺伝子情報推定方法に関する。
【００１３】
ステップ（ｉｖ）、（ｖ）、および（ｖｉｉ）において、過分散の程度を表すパラメータをϕ_iとし、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数をμ_{ski}としたとき、サンプル（ｓ）における遺伝子（ｉ）の出現数（r_{si}）が、
【数２】

$$P(r_{si}) = \frac{\Gamma\left(r_{si} + \phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)} \left(\frac{\mu_{ski}}{\mu_{ski} + \phi_i^{-1}}\right)^{r_{si}} \left(\frac{1}{1 + \mu_{ski}\phi_i}\right)^{\phi_i^{-1}}$$

という確率に従うことが好ましい。
【００１４】
サンプル（ｓ）における全遺伝子のカウントの総和（ν_s）、全サンプルでの全遺伝子の出現数の総和に対する遺伝子（ｉ）の出現数の割合の対数（m_i）、指数分布あるいはジェフリーズ事前分布に従う分散（τ_{ki}）、および平均が０で分散がτ_{ki}の正規分布に従うトピック（ｋ）における遺伝子（ｉ）の期待カウントの平均からのずれを表す量が与えられたときに、サンプル（ｓ）においてトピック（ｋ）を割り当てられた遺伝子（ｉ）の期待出現数（μ_{ski}）が
$\mu_{ski} = \nu_s \exp(\eta_{ki} + m_i)$
と表される。
【発明の効果】
【００１５】
過分散のある遺伝子情報を扱うための新しいトピックモデルを提案し、シミュレーションによって推定精度の検証を行った。低コストを想定した総リード数が10^6程度のデータから本発明のモデルによって推定した遺伝子発現量は、よりコストのかかる総リード数10^9程度で定量した発現量よりも真の値に近かった（図８）。このことは、１サンプル当たりの総リード数を増やすことで質の良いデータを得るよりも、総リード数を減らして多くのサンプルの定量を行ったうえで本発明のモデルによる発現量の推定を行うことが有効であることを示している。
【図面の簡単な説明】

【００１６】
【図１】本発明の実施形態に係る遺伝子情報推定装置１の構成を示すブロック図を示す。
【図２】遺伝子発現推定装置１による処理の流れを示すフローチャートを示す。
【図３Ａ】トピックモデル推定プログラム４１の一実施形態による処理の流れを示すフローチャートを示す。
【図３Ｂ】トピックモデル推定プログラム４１の他の実施形態による処理の流れを示すフローチャートを示す。
【図４】遺伝子情報推定プログラム４３による処理の流れを示すフローチャートを示す。
【図５】実施例における対数周辺尤度の下限の近似値を示す。
【図６】実施例において推定された各サンプルにおけるトピックの出現確率を示す。
【図７】実施例における各遺伝子の出現確率のトピック間の差を示す。
【図８】実施例におけるリード数の期待値と推定値との誤差の分布を示す。
【発明を実施するための形態】
【００１７】
［１］遺伝子情報推定装置
本発明は、
（ｉ）複数のサンプルに対応する遺伝子情報の測定データを読み込む手順、
（ｉｉ）遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出する手順、
（ｉｉｉ）各サンプルの各遺伝子にトピックが割り当てられる確率を初期化する手順、
（ｉｖ）遺伝子へのトピック割り当て確率を更新する手順、
（ｖ）トピック情報を更新する手順、
（ｖｉ）トピック情報の分散を更新する手順、
（ｖｉｉ）負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新する手順、
および、
（ｖｉｉｉ）トピック情報およびトピック出現頻度から遺伝子情報を推定する手順、
を有する、
トピックモデルを用いた遺伝子情報推定装置に関する。
【００１８】
手順（ｉ）では複数のサンプルに対応する遺伝子情報の測定データを読み込む。手順（ｉｉ）では遺伝子情報の測定データに基づき遺伝子の平均出現頻度を算出する。遺伝子の平均出現頻度の算出法については後述する。手順（ｉｉｉ）では各サンプルの各遺伝子にトピックが割り当てられる確率を初期化する。初期化方法については後述する。
【００１９】
遺伝子へのトピック割り当て確率を更新する手順（ｉｖ）では、各サンプルにおいて各遺伝子のカウントがどのトピックから生成されたかについて事後分布の推定を行う。
【００２０】
トピック情報を更新する手順（ｖ）では、割り当てられたトピックに基づき、トピック情報、すなわち各トピックにおける遺伝子の出現頻度の平均からのずれを推定する。
【００２１】
トピック情報の分散を更新する手順（ｖｉ）では、推定されたトピック情報に基づき、トピック情報の分散を更新する。
【００２２】
負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを更新する手順（ｖｉｉ）では、トピックごとに各遺伝子のカウントデータを生成する負の二項分布のパラメータを推定する。すなわち、手順（ｉｖ）～（ｖ）で割り当てられたトピックと推定されたトピック情報から算出される期待値と測定データの差分に基づき、負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータを推定する。
【００２３】
トピック情報およびトピック出現頻度から遺伝子情報を推定する手順（ｖｉｉｉ）では、トピックモデル推定プログラムにより推定されたトピック情報と平均出現頻度から各トピ

ックにおける遺伝子の出現頻度を算出し、同じくトピックモデル推定プログラムで推定された各サンプルにおけるトピックの出現頻度と掛け合わせて得られる各サンプルにおける遺伝子の出現頻度を遺伝子情報の推定として出力する。

【００２４】

［２］遺伝子情報の測定データ

遺伝子情報の測定データは、試料から核酸を調製する工程、核酸からライブラリーを調製し増幅する工程、および次世代シークエンサーによりライブラリーの配列決定を行う工程を経ることにより得られる。各工程における作業手順および条件は特に限定されず、通常の方法を用いることが可能である。試料は限定されず、植物細胞、微生物細胞、動物細胞等を使用可能である。核酸はＤＮＡであってもよく、ＲＮＡであってもよい。遺伝子情報は、複数の遺伝子の発現や、複数の遺伝子の存在に関連する情報であれば特に限定されないが、本発明の遺伝子情報推定装置は、過分散を持つ遺伝子情報に好適に適用できる。

【００２５】

遺伝子情報の測定データとして、ＲＮＡ－Ｓｅｑデータ、ＣｈＩＰ－Ｓｅｑデータが挙げられる。ＲＮＡ－Ｓｅｑデータは、ＲＮＡを次世代シークエンサーで定性・定量して得られ、トランスクリプトーム解析に用いられる。ＣｈＩＰ－Ｓｅｑデータは、クロマチン免疫沈降法と次世代シークエンサーによる配列決定を組み合わせて得られ、ＤＮＡ結合タンパク質が結合するＤＮＡ配列を示す。ＣｈＩＰ－Ｓｅｑデータの取得時に対象となるＤＮＡ結合タンパク質としては、ヒストン、転写因子、ＤＮＡ修飾酵素が挙げられる。

【００２６】

遺伝子情報の測定データとして、ＤＮＡのメチル化を示すデータも挙げられる。ＤＮＡのメチル化を示すデータは、バイサルファイト変換を行ったゲノムＤＮＡを次世代シークエンサーで解析してメチル化シトシンを検出する方法や、抗メチルシトシン抗体やメチルＣｐＧタンパク質によるＤＮＡの沈降と次世代シークエンサーによる配列決定を組み合わせる方法により得ることができる。

【００２７】

本発明の装置は、メタゲノム解析において遺伝子情報を推定するためにも使用可能である。メタゲノム解析は、複数の微生物を含む試料から核酸を抽出し、その核酸塩基配列をまとめて解読し、試料中に存在した微生物の構成を解析する手法である。

【００２８】

遺伝子情報の測定データとして、ＲＮＡ－Ｓｅｑデータを用いた場合、出力される遺伝子情報は、リード数、すなわち遺伝子発現量の期待値である。これにより、試料におけるトランスクリプトームを高い精度で推定できる。

【００２９】

通常、遺伝子情報の推定のためにはサンプルあたりのデータ量が多い必要があるが、本発明の装置によれば、ＲＮＡ－Ｓｅｑの場合にはサンプルあたり１００，０００から１０００，０００リード程度という少ないデータ量でも高い精度で遺伝子情報を推定できる。

【００３０】

［３］遺伝子情報推定装置の全体構成

図１は、本発明の実施形態に係る遺伝子情報推定装置１の構成を示すブロック図である。遺伝子情報推定装置１は、演算手段２と、入力手段３と、記憶手段４と、出力手段５とを備えている。各手段２～５はバスライン６に接続されている。

【００３１】

演算手段２は、例えば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）およびＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）から構成される主制御装置である。演算手段２は、トピックモデル推定部２１と、遺伝子情報推定部２２と、メモリ２３を含む。

【００３２】

入力手段３は、例えば、キーボード、マウス、ディスクドライブ装置などから構成される。記憶手段４は、例えば、一般的なハードディスク装置などから構成され、プログラム格

納部４０と、データ格納部４４とを含む。
【００３３】
プログラム格納部４０には、演算手段２で用いられるプログラムとして、トピックモデル推定プログラム４１、および遺伝子情報推定プログラム４３とを記憶させておくことが可能である。演算手段２は、記憶手段４のプログラム格納部４０から、トピックモデル推定プログラム４１、および遺伝子情報推定プログラム４３をそれぞれ読み込み、メモリ２３に格納し実行することで、トピックモデル推定部２１と、遺伝子情報推定部２２とをそれぞれ実現する。
【００３４】
データ格納部４４には、演算手段２で用いられる各種データとして、遺伝子情報の測定データ４５、遺伝子の平均出現頻度４６、トピック情報４７、トピック出現頻度４８とを記憶する。遺伝子情報の測定データ４５は入力手段３を介して入力され、記憶手段４のデータ格納部４４に記憶される構成とすることが可能である。遺伝子の平均出現頻度４６、トピック情報４７、トピック出現頻度４８は、演算手段２の演算処理結果を示すデータである。
【００３５】
出力手段５は、例えば、グラフィックボード（出力インタフェース）およびそれに接続されたモニタである。モニタは、例えば、液晶ディスプレイ等から構成され、可視化を行った結果等を表示する。
【００３６】
本発明の課題は、前述の遺伝子情報推定装置を構成する各手段としてコンピュータを機能させるための遺伝子情報推定プログラム、前記遺伝子情報推定プログラムが記録されたコンピュータ読み取り可能な記録媒体によっても解決できる。また、本発明の課題は、前述の遺伝子情報推定装置の手順と同様のステップ含む、トピックモデルを用いた遺伝子情報推定方法によっても解決できる。
【００３７】
遺伝子情報推定装置は、一般的なコンピュータを、遺伝子情報推定プログラムにより動作させることで実現できる。このプログラムは、通信回線を介して提供することも可能であるし、ＣＤ－ＲＯＭ等のコンピュータ読み取り可能な記録媒体に書き込んで配布することも可能である。コンピュータは、コンピュータ読み取り可能な記録媒体から遺伝子情報推定プログラムを読み出し、このプログラムに基づいた各機能を実現することができる。また、遺伝子情報推定プログラムをインストールされたコンピュータは、ＣＰＵが、ＲＯＭ等に格納されたこのプログラムをＲＡＭに展開することにより、遺伝子情報推定装置としての効果を奏することができる。
【００３８】
［４］遺伝子情報推定装置による処理の流れ
図２は、図１に示した遺伝子情報推定装置１による処理の流れを示すフローチャートである。Ｓ１１において、遺伝子情報推定装置１は遺伝子情報の測定データ４５を読み込む。Ｓ１２において、遺伝子情報推定装置１は遺伝子情報の測定データ４５から遺伝子の平均出現頻度４６を算出する。
【００３９】
遺伝子の平均出現頻度４６は、例えば遺伝子の全サンプルを通しての出現回数を、その全ての遺伝子についての総和で割ることで算出できる。
【００４０】
Ｓ１３において、遺伝子情報推定装置１はトピックモデル推定プログラム４１によりトピック情報４７およびトピック出現頻度４８を推定する。Ｓ１４において、遺伝子情報推定装置１はトピック情報４７およびトピック出現頻度４８から遺伝子情報を推定する。Ｓ１５において、遺伝子情報推定装置１は推定された遺伝子情報を出力手段５から出力する。
【００４１】
［５］トピックモデル推定プログラムによる処理の流れ

図３Ａは、図１に示したトピックモデル推定プログラム４１による処理の流れを示すフローチャートである。

【００４２】

Ｓ２０１において、遺伝子情報推定装置１は、トピックモデル推定プログラム４１の初期化を行う。Ｓ２０２において、遺伝子情報推定装置１は遺伝子へのトピック割り当て確率の更新を行う。Ｓ２０３において、遺伝子情報推定装置１はトピック情報（η）の更新１を行う。Ｓ２０４において、遺伝子情報推定装置１はトピック情報の分散（τ）の更新を行う。Ｓ２０５において、遺伝子情報推定装置１は負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータ（φ）の更新を行う。Ｓ２０６において、遺伝子情報推定装置１は対数周辺尤度の下限が収束しているか否かを評価し、収束していない場合にはＳ２０２に戻る。対数周辺尤度の下限が収束している場合にはＳ２１１に進む。Ｓ２１１において、遺伝子情報推定装置１はトピック情報４７およびトピック出現頻度４８を出力手段５から出力する。

【００４３】

以下、トピックモデル推定プログラムにおける処理の流れの詳細を、必要に応じて図３Ａを参照しながら説明する。過分散なデータを単一のトピックから発生させるために、先行技術であるＳｐａｒｓｅ Ａｄｄｉｔｉｖｅ Ｇｅｎｅｒａｔｉｖｅ ｍｏｄｅｌ（ＳＡＧＥ；Ｅｉｓｅｎｓｔｅｉｎ ｅｔ ａｌ．Ｐｒｏｃ．Ｉｎｔ．Ｃｏｎｆ．Ｍａｃｈ．Ｌｅａｒｎ．２０１１）における単語あるいは遺伝子の出現確率を負の二項分布に置き換える。サンプルで各遺伝子に割り当てられるトピックの確率は、ＳＡＧＥと同様にＤｉｒｉｃｈｌｅｔ分布から生成される。

【００４４】

【数３】

$$\boldsymbol{\theta}_s \sim \mathcal{D}ir(\boldsymbol{\alpha}),\ s = 1, \ldots, S$$

【数４】

$$z_{si} \sim \mathcal{M}ulti(\boldsymbol{\theta}_s),\ i = 1, \ldots, G$$

【００４５】

ここで、θ_sはサンプルｓにおいて各トピックを割り当てる確率、z_{si}はサンプルｓのｉ番目の遺伝子に割り当てられたトピック、ｓとＧはそれぞれサンプルと遺伝子の数を表す。先行技術と同様に、すべてのトピックにおける平均的な遺伝子の出現頻度の対数をｍ、各トピックにおける出現頻度がこのｍからどの程度ずれているかを表す量をη_kとし、その事前分布は以下の正規分布とする。

【００４６】

【数５】

$$\eta_{k,i} \sim \mathcal{N}(0, \tau_{k,i})$$

【００４７】

分散τ_{ki}は指数分布

【数６】

$$P(\tau_{k,i}) \propto \exp(-\gamma\tau_{ki})$$

あるいはＪｅｆｆｒｅｙｓ事前分布

【数７】

$$P(\tau_{k,i}) \propto 1/\tau_{k,i}$$

を用いる。こうすることで、η_kは多くの要素が０となることが知られている。

サンプルｓにおける遺伝子ｉにトピックｋが割り当てられたとき、すなわちz_{si}＝ｋのとき、そのリード数は、

【数８】

$$P(r_{si}) = \frac{\Gamma\left(r_{si} + \phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)} \left(\frac{\mu_{ski}}{\mu_{ski} + \phi_i^{-1}}\right)^{r_{si}} \left(\frac{1}{1 + \mu_{ski}\phi_i}\right)^{\phi_i^{-1}}$$

という負の二項分布に従うものとする。ここで、φ$_{i}$は負の二項分布において過分散の度合いを決めるｄｉｓｐｅｒｓｉｏｎというパラメータである。この分布の期待出現数μ$_{sk\,i}$は、サンプルｓにおける全遺伝子のカウントの総和をν$_{s}$と表記すると、

【数９】

$$\mu_{ski} = \nu_s \lambda_{k,i} = \nu_s \exp\left(m_i + \eta_{k,i}\right)$$

となる。

【００４８】

トピックモデル推定プログラムの初期化処理（Ｓ２０１）では、各サンプルの各遺伝子にトピックの割り当てられる確率をランダムに初期化する。また、入力されたカウントデータから平均的な出現頻度の対数ｍ$_{i}$を算出し、上記のトピックの割り当てられる確率と合わせて平均出現頻度からのずれη$_{k\,i}$の初期値を決める。

【００４９】

遺伝子へのトピック割り当て確率の更新（Ｓ２０２）は下記の手順により行う。すなわち、ＲＮＡ－Ｓｅｑデータに基づき、ｚ、η、τの事後分布の推定を行う。これらの事後分布を解析的に求めることはできないため、事後分布の近似ｑ（ｚ）、ｑ（η）、ｑ（τ）を考え、対数周辺尤度の下限

【数１０】

$$\begin{aligned} F\left(q(\boldsymbol{z}), q(\boldsymbol{\eta}), q(\boldsymbol{\tau})\right) =& \mathbb{E}_{q(\boldsymbol{z}),q(\boldsymbol{\eta})}\left[\log P(\boldsymbol{r}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\phi}) + \log P(\boldsymbol{\eta}|\boldsymbol{\tau}) + \log P(\boldsymbol{\tau})\right] \\ & - \mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z})\right] - \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log q(\boldsymbol{\eta})\right] - \mathbb{E}_{q(\boldsymbol{\tau})}\left[\log q(\boldsymbol{\tau})\right] \end{aligned}$$

を最大化する。ここで

【数１１】

$$\mathbb{E}_q[\cdot]$$

は確率分布ｑ（）のもとでの期待値を意味する。同様に確率分布ｑ（）のもとでの分散を

【数１２】

$$\mathbb{V}_q[\cdot]$$

、サンプルｓの遺伝子ｉを除いたｚをｚ$^{\backslash s,\,i}$、サンプルｓにおいて遺伝子ｉ以外でトピックｋを割り当てられた遺伝子の数をｎ$_{s,\,k}{}^{\backslash s,\,i}$と表記し、

【数１３】

$$\hat{\lambda}_{k,i} = \exp\left(m_i + \mathbb{E}_{q(\boldsymbol{\eta})}\left[\eta_{k,i}\right]\right)$$

とおく。

【００５０】

トピックモデルの推定は、τ$_{k\,i}$の事前分布を指数分布として推定を行う。最初に遺伝子へのトピックの割り当て確率の更新（Ｓ２０２）を行う。トピックの割り当て確率ｑ（ｚ）の更新式は、

【数１４】

$$\log q(z_{si}=k)$$

$$\approx \log\left(\mathbb{E}_{q(\boldsymbol{z}^{\setminus s,i})}\left[n_{s,k}^{\setminus s,i}\right]+\alpha_k\right)-\frac{\mathbb{V}_{q(\boldsymbol{z}^{\setminus s,i})}\left[n_{s,k}^{\setminus s,i}\right]}{2\left(\mathbb{E}_{q(\boldsymbol{z}^{\setminus s,i})}\left[n_{s,i}^{\setminus s,i}\right]+\alpha_k\right)^2}$$
$$-r_{si}\log\left(1+1/\nu_s\hat{\lambda}_{k,i}\phi_i\right)-\phi_i^{-1}\log\left(1+\nu_s\hat{\lambda}_{k,i}\phi_i\right)$$
$$-\frac{\left(r_{si}+\phi_i^{-1}\right)\mathbb{V}_{q(\boldsymbol{\eta})}\left[\eta_{k,i}\right]}{2\left(1+1/\nu_s\hat{\lambda}_{k,i}\phi_i\right)\left(1+\nu_s\hat{\lambda}_{k,i}\phi_i\right)}+\text{const.}$$

となる。

【００５１】

次に、トピック情報（η）の更新１（Ｓ２０３）を行う。ｑ（η）はラプラス近似を行う。関数

【数１５】

$$f(\eta_{k,i})=\mathbb{E}_{q(\boldsymbol{z}),q(\boldsymbol{\eta})}\left[-r_{si}\log\left(1+\frac{1}{\nu_s\lambda_{k,i}\phi_i}\right)-\phi_i^{-1}\log\left(1+\nu_s\lambda_{k,i}\phi\right)\right]-\frac{\mathbb{E}_{q(\boldsymbol{\tau})}\left[\tau_{k,i}^{-1}\right]\mathbb{E}_{q(\boldsymbol{\eta})}\left[\eta_{k,i}^2\right]}{2}$$

を定義し、

【数１６】

$$\eta_{k,i}^*=\arg\max_{\eta_{k,i}} f(\eta_{k,i})$$

と表記すると

【数１７】

$$q(\eta_{k,i})\approx\mathcal{N}\left(\eta_{k,i}^*,-\frac{1}{f''(\eta_{k,i}^*)}\right)$$

となる。ここで、

【数１８】

$$\eta_{k,i}^*$$

の推定値はニュートン法による更新を１ステップにつき１回行う、

【数１９】

$$\Delta\eta_{k,i}^*=-\frac{f'(\eta_{k,i}^*)}{f''(\eta_{k,i}^*)}$$

【００５２】

次に、トピック情報の分散（τ）の更新（Ｓ２０４）を行う。τの事後分布ｑ（τ）はガンマ分布

【数２０】

$$q(\tau_{k,i}|a_{k,i},b_{k,i})=\tau_{k,i}^{a_{k,i}-1}\frac{\exp(-\tau_{k,i}/b_{k,i})}{\Gamma(a_{k,i})b_{k,i}^{a_{k,i}}}$$

とし、そのパラメータａｋ，ｉとｂｋ，ｉを

【数２１】

$$-\Delta a_{k,i}=\frac{(\frac{1}{2}-a_{k,i})\psi_1(a_{k,i})+\frac{\mathbb{E}_{q(\boldsymbol{\eta})}\left[\eta_{k,i}^2\right]}{2(a_{k,i}-1)^2 b_{k,i}}-\gamma b_{k,i}+1}{(\frac{1}{2}-a_{k,i})\psi_2(a_{k,i})-\psi_1(a_{k,i})-\frac{\mathbb{E}_{q(\boldsymbol{\eta})}\left[\eta_{k,i}^2\right]}{(a_{k,i}-1)^3 b_{k,i}}}$$

【数２２】

$$b_{k,i}=\frac{1+\sqrt{1+8\gamma\mathbb{E}_{q(\boldsymbol{\eta})}\left[\eta^2\right]a_{k,i}/(a_{k,i}-1)}}{4\gamma a_{k,i}}$$

で更新する。

【００５３】

以上の推定値から算出される期待出現頻度と測定されたカウントデータｒ$_{si}$との差分に基づき、負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータφの更新を行う（Ｓ２０５）。以上の更新（Ｓ２０２～Ｓ２０５）を対数周辺尤度の下限が収束するまで繰り返す（Ｓ２０６）。対数周辺尤度の下限は各更新ステップで導出された近似に基づき算出される。

【００５４】

より高精度の推定を行うために、トピックモデルの推定は複数段階行ってもよい。トピックモデルの推定を二段階行う場合のトピック推定プログラム４１による処理の流れを図３Ｂに示す。

【００５５】

図３Ｂにおいて、Ｓ２０１～Ｓ２０６およびＳ２１１は図３Ａと同じである。Ｓ２０６において、遺伝子情報推定装置１は対数周辺尤度の下限が収束しているか否かを評価し、収束していない場合にはＳ２０２に戻る。対数周辺尤度の下限が収束している場合にはＳ２０７に進む。Ｓ２０７において、遺伝子情報推定装置１は遺伝子へのトピック割り当て確率の更新を行う。Ｓ２０８において、遺伝子情報推定装置１はトピック情報（η）の更新２を行う。Ｓ２０９において、遺伝子情報推定装置１は、負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータ（φ）の更新を行う。Ｓ２１０において、遺伝子情報推定装置１は対数周辺尤度の下限が収束しているか否かを評価し、収束していない場合にはＳ２０７に戻る。対数周辺尤度の下限が収束している場合にはＳ２１１に進む。Ｓ２１１において、遺伝子情報推定装置１はトピック情報４７およびトピック出現頻度４８を出力手段５から出力する。

【００５６】

トピックモデルの推定を二段階行う場合、まず、τ$_{ki}$の事前分布を指数分布として前述の方法で推定を行う。対数周辺尤度の下限が収束した後、よりスパースなηを得るために、τ$_{ki}$の事前分布をＪｅｆｆｒｅｙｓ事前分布として遺伝子へのトピック割り当て確率（Ｓ２０７）、トピック情報（Ｓ２０８）、負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータ（Ｓ２０９）の更新を、対数周辺尤度の下限がもう一度収束するまで繰り返す。ｑ（ｚ）は一段階目と同じ式で更新できる。また、ｑ（τ）はキャンセルされて消えるため、二段階目では更新を行わない。ｑ（η）の更新のみ、

【数２３】

$$f'(\eta_{k,i})$$

$$f''(\eta_{k,i})$$

中の

【数２４】

$$\mathbb{E}_{q(\boldsymbol{\tau})}\left[\tau_{k,i}^{-1}\right]$$

が、

【数２５】

$1/\eta_{k,i}$

となるために更新式が異なる（Ｓ２０８）。

【００５７】

遺伝子へのトピック割り当て確率の更新、トピック情報（η）の更新２、および負の二項分布のｄｉｓｐｅｒｓｉｏｎパラメータの更新（Ｓ２０７～Ｓ２０９）を、対数周辺尤度の下限が収束するまで繰り返す（Ｓ２１０）。対数周辺尤度の下限の収束は、Ｓ２０６と同じ方法により評価できる。対数周辺尤度の下限が収束した後、トピック情報４７およびトピック出現頻度４８が出力される（Ｓ２１１）。

【００５８】

Ｓ２０２～Ｓ２０５、およびＳ２０７～Ｓ２０９において、ｄｉｓｐｅｒｓｉｏｎφは負の二項分布の分散と期待値とｄｉｓｐｅｒｓｉｏｎの関係、

【数２６】

$$\mathbb{V}[r_{s,i}] = \mathbb{E}[r_{s,i}] + \mathbb{E}[r_{s,i}]^2 \phi_i$$

から、リード数の期待値と実測の二乗誤差に基づいて推定する。

【００５９】

［５］遺伝子情報推定プログラムによる処理の流れ

図４は、図１に示した遺伝子情報推定プログラム４３による処理の流れを示すフローチャートである。Ｓ３１において、遺伝子情報推定装置１は、遺伝子の平均出現頻度４６、トピック情報４７、およびトピック出現頻度４８を読み込む。遺伝子の平均出現頻度４６は図２のＳ１２で算出されたものであり、トピック情報４７、およびトピック出現頻度４８は図３Ａおよび図３ＢのＳ２１１で出力されたものである。Ｓ３２において、遺伝子情報推定装置１は遺伝子情報を推定する。Ｓ３３において、遺伝子情報推定装置１は、推定された遺伝子情報を出力手段５から出力する。

【００６０】

Ｓ３２における遺伝子情報の推定において、サンプル（ｓ）における遺伝子（ｉ）の期待出現数

【数２７】

$\hat{\mu}_{si}$

は、

サンプル（ｓ）における全遺伝子のカウントの総和（ν_s）、トピックモデル推定プログラム（Ｓ１３）により推定された全サンプルでの全遺伝子の出現数の総和に対する遺伝子（ｉ）の出現数の割合の対数（m_i）、および、以上で推定された各サンプルにおける各遺伝子へのトピックの割り当て確率、トピック（ｋ）における遺伝子（ｉ）の期待カウントの平均からのずれを表す量（η）から、

【数２８】

$$\hat{\mu}_{si} = \sum_k q(z_{ki}) \mathbb{E}_{q(\eta)}[\lambda_{k,i}] = \sum_k q(z_{s,k}) \exp\left(m_i + \eta^*_{k,i} + \frac{\mathbb{V}_{q(\eta)}[\eta_{k,i}]}{2}\right)$$

と推定される。

【実施例】

【００６１】

(実施例１)トピックの出現確率の推定

先行技術との比較を行うために、疑似的なＲＮＡ－Ｓｅｑデータを作成して解析を行った。統制群と実験群がそれぞれ２０サンプルある状況を想定した。遺伝子の総数を１０，０００とし、そのうち５００遺伝子は実験群で発現量が増加し、別の５００遺伝子は発現量が減少する。残りの９，０００遺伝子は二群間で発現量が変わらないとする。統制群にお

ける各遺伝子のリード数は以下の負の二項分布に従う。

【００６２】

【数２９】

$$\log \bar{\mu}_i \sim \mathcal{N}(0,1)$$

【数３０】

$$\mu_i^c = \frac{\bar{\nu}\bar{\mu}_i}{\sum_i \bar{\mu}_i}$$

【数３１】

$$P(r_{si}) = \frac{\Gamma\left(r_{si}+\phi_i^{-1}\right)}{r_{si}!\Gamma\left(\phi_i^{-1}\right)}\left(\frac{\mu_i^c}{\mu_i^c+\phi_i^{-1}}\right)^{r_{si}}\left(\frac{1}{1+\mu_i^c\phi_i}\right)^{\phi_i^{-1}}$$

【００６３】

【数３２】

$\bar{\nu}$

は総リード数の期待値で、低コストで総リード数の少ない条件

【数３３】

$$\bar{\nu} = 10^6$$

と高コストで総リード数の多い条件

【数３４】

$$\bar{\nu} = 10^8$$

の二条件を想定した。ｄｉｓｐｅｒｓｉｏｎφｉは、公開されている実測のデータ（Ｐｉｃｋｅｒｅｌｌ　ｅｔ　ａｌ．Ｎａｔｕｒｅ　２０１０）に基づきｅｄｇｅＲ（Ｚｈｏｕ　ｅｔ　ａｌ．Ｎｕｃｌｅｉｃ　Ａｃｉｄｓ　Ｒｅｓ．２０１４）で推定を行った。実験群の期待リード数μｉ^ｅは以下のように決定した。

【００６４】

【数３５】

$$\mu_i^e = \mu_i^c\,(\gamma_i + 1.5)^{\rho_i}$$

【数３６】

$$P(\gamma_i) \propto \exp(-\gamma_i)$$

【００６５】

ここで、ρｉは実験群で発現量が増加する遺伝子群については、ρｉ＝＋１、減少する遺伝子群についてはρｉ＝－１、残りの変化しない遺伝子群については、ρｉ＝０とした。生成した疑似データについて、ＬＤＡ、ＳＡＧＥ、本発明の提案モデルの３つのモデルを適用した。

【００６６】

異なるトピック数で提案モデルの推定を行った時の、対数周辺尤度の下限の近似値を図５に示す。図５の横軸はトピック数を表し、縦軸は変分下限の近似値を表す。トピック数が２の時に最大となることから、以降の解析ではトピック数を２とした。疑似データは統制群と実験群の二群からなるため、この結果はトピック数を適切に選択できていることを示している。

【００６７】

３つのモデルで推定された各サンプルにおけるトピックの出現確率を図６に示す。図６において、横軸はサンプルを表し、縦軸はトピックの出現確率を表す。統制群と実験群を左

右に分けて表示している。すべてのモデルで、二群間でトピックの出現確率が変化しているが、ＬＤＡとＳＡＧＥに対して、本発明のモデル（ｐｒｏｐｏｓｅｄ）では二群間の違いが顕著になっている。

【００６８】

(実施例２)トピックにおける遺伝子の出現確率の比較

各遺伝子について推定された二つのトピックにおける出現確率の差を比較した。ＬＤＡについては、各トピックにおける出現確率の差を、ＳＡＧＥと本発明のモデルについては、

【数３７】

$$\eta_{2,i} - \eta_{1,i}$$

を計算し、ヒストグラムを作成した（図７）。図７において、実験群で発現量が増加する５００遺伝子（ｕｐ）を白色、減少する５００遺伝子（ｄｏｗｎ）を黒色、それ以外の９，０００遺伝子（ｃｏｎｓｔ）を灰色で表示している。ＬＤＡでは、ほとんどの遺伝子が０付近に集中しており、発現量の増加する遺伝子と減少する遺伝子とそれ以外の遺伝子を区別することが困難である。３種類の遺伝子は、ＳＡＧＥよりも本発明のモデル（ｐｒｏｐｏｓｅｄ）の方がより明瞭に分離している。

【００６９】

(実施例３) 発現量の推定精度の比較

発現量の推定精度を比較するために、リード数の期待値μ_{si}と推定値の誤差を算出した（図８）。図８ではリード数の期待値の大きさに応じて分割してプロットした。期待総リード数10^6で生成した疑似データ（ｒａｗ）、コストをかけた状態を想定して期待総リード数を10^8とした疑似データ（ｄｅｅｐ）、ＳＡＧＥの推定（ｓａｇｅ）、本発明のモデルの推定（ｐｒｏｐｏｓｅｄ）を、それぞれ黒色の実線、灰色の実線、黒色の破線、灰色の破線でプロットした。

【００７０】

総リード数の多い状況を想定したデータ（ｄｅｅｐ）は総リード数をそろえるために、生成したデータを１００で割った値と期待値を比較した。ＬＤＡは、元のデータ（ｒａｗ）よりも誤差が大きくなっており、発現量の推定には使用できない。ｓａｇｅと本発明のモデルを比較すると、約６０％の観測で本発明のモデルの方が期待値に近い推定を与えている。

【００７１】

本発明のモデルは、今回のシミュレーションの二群比較のように統制のとれた実験のサンプルだけではなく、ｅｄｇｅＲやＤＥＳｅｑといった負の二項分布を用いた既存の解析法が利用できない未知の構成のサンプルから定量したＲＮＡ－Ｓｅｑデータに対しても、有効と考えられる。

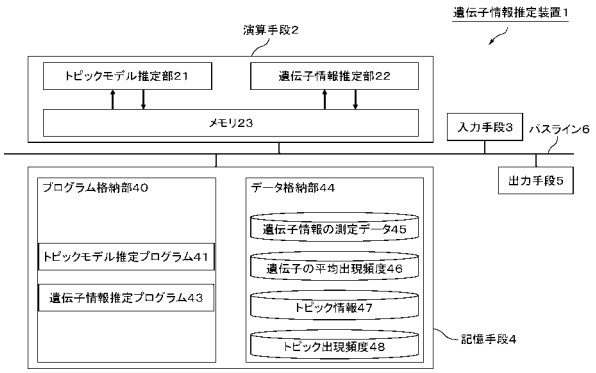
【符号の説明】

【００７２】

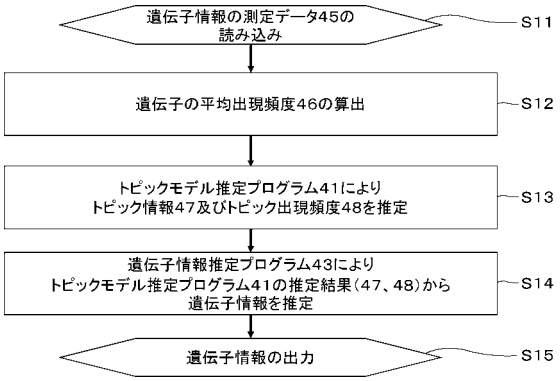
１　遺伝子情報推定装置
２　演算手段
２１　トピックモデル推定部
２２　遺伝子情報推定部
２３　メモリ
３　入力手段
４　記憶手段
４０　プログラム格納部
４１　トピックモデル推定プログラム
４３　遺伝子情報推定プログラム
４４　データ格納部
４５　遺伝子情報の測定データ
４６　遺伝子の平均出現頻度

４７　トピック情報
４８　トピック出現頻度
５　　出力手段
６　　バスライン

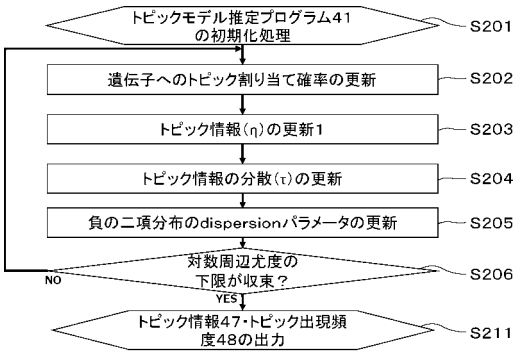
【図１】



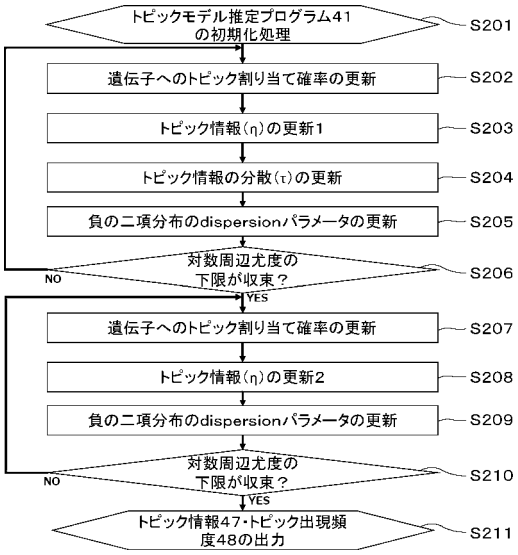
【図２】



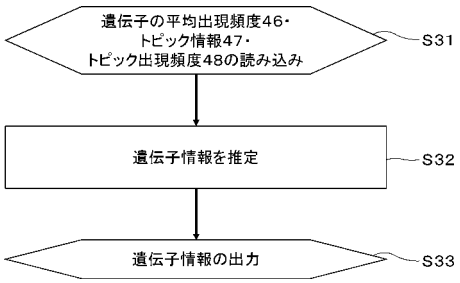
【図３Ａ】



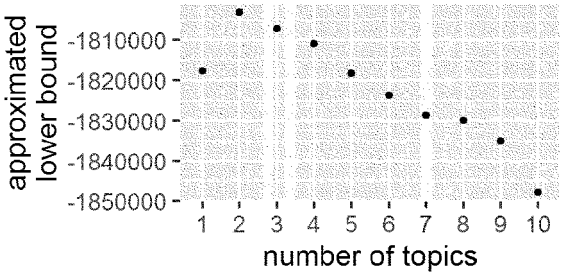
【図３Ｂ】



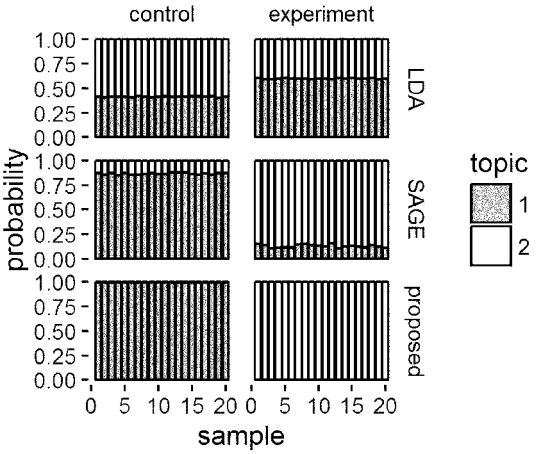
【図４】



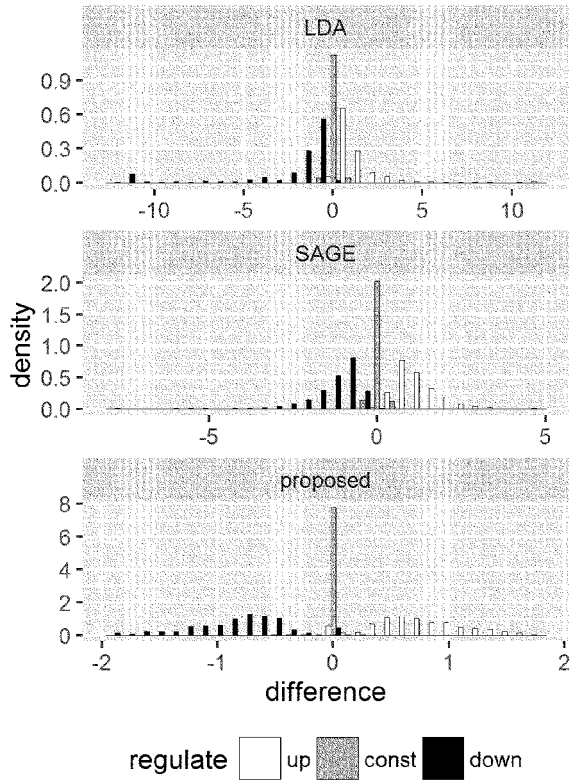
【図５】



【図６】



【図７】



【図８】